2007

# Schema-Driven Relationship Extraction from Unstructured Text

Cartic Ramakrishnan
*Wright State University - Main Campus*

# Schema-Driven Relationship Extraction from Unstructured Text

Cartic Ramakrishnan
Kno.e.sis Center, Wright State University,
Dayton, OH

# Outline

Motivation

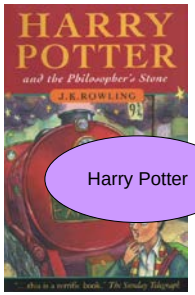
Problem Description & Approach

Results

Future Work

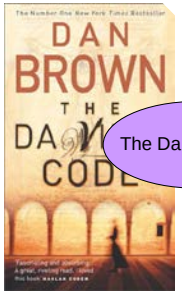
# Anecdotal Example

UNDISCOVERED PUBLIC KNOWLEDGE
Discovering connections hidden in text




Harry Potter

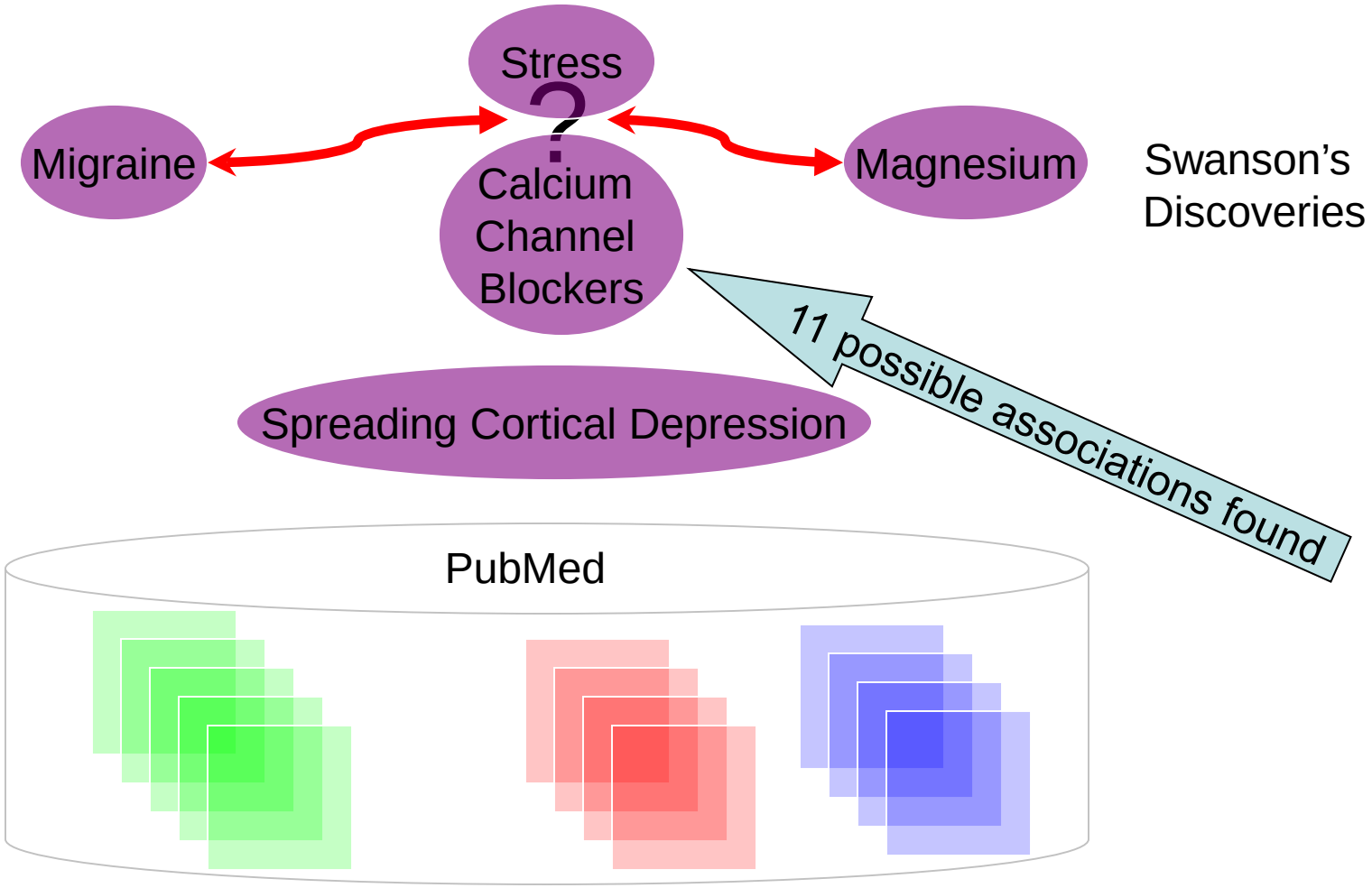
Et in Arcadia Ego




The Da Vinci code

Santa Maria delle Grazie

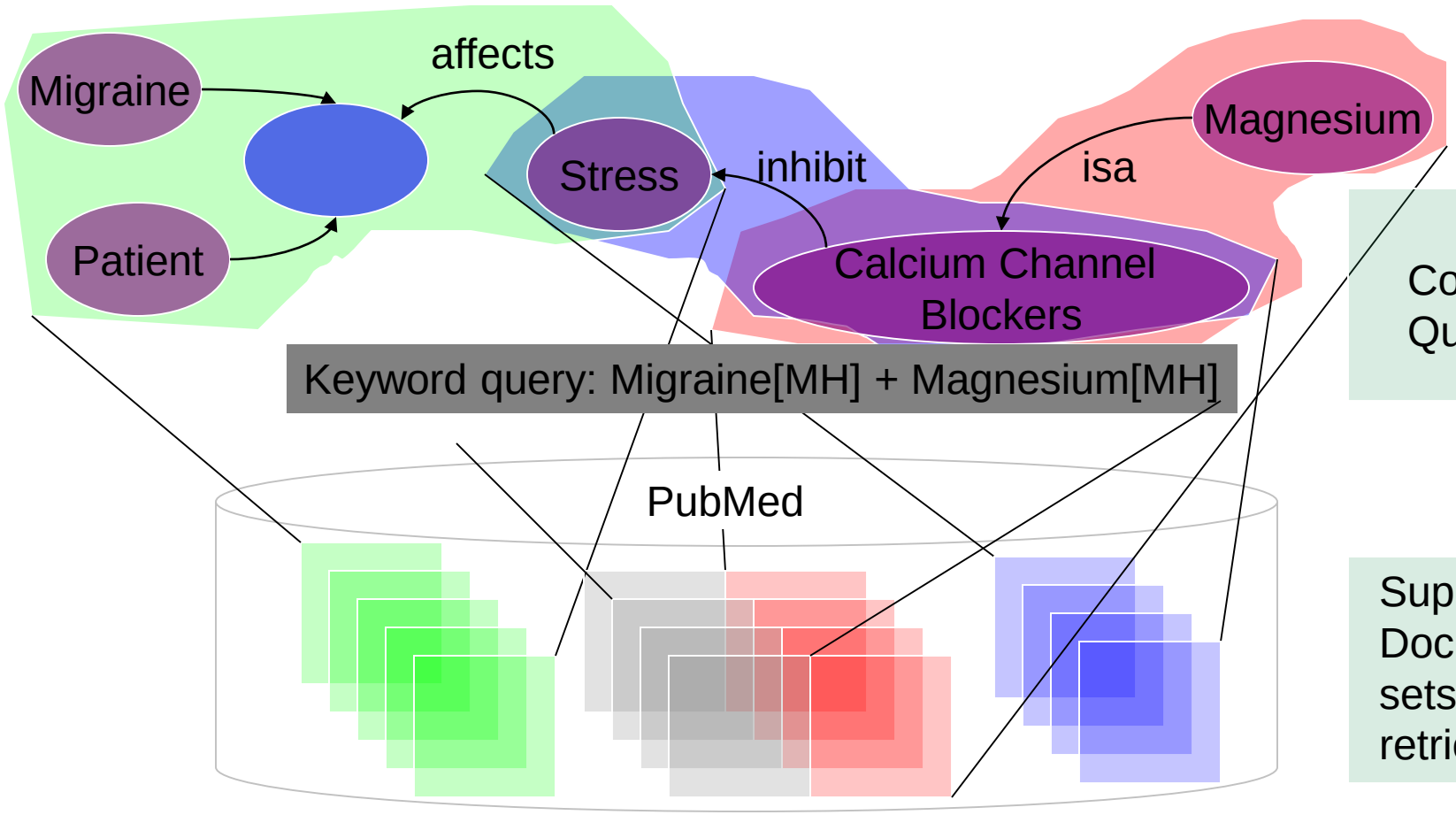
# Motivation 1 – Undiscovered Public knowledge in biology




Associations Discovered based on keyword searches followed by manually analysis of text to establish possible relevant relationships

These associations were discovered in 1986

# Motivation 2 - Hypothesis Driven retrieval of Scientific Literature



# Motivation 3 -- Growth Rate of Public Knowledge

Data captured per year = 1 exabyte ($10^{18}$)
(Eric Neumann, Science, 2005)

How much is that?

- Compare it to the estimate of the total words ever spoken by humans = 12 exabyte

A small but significant portion is text data

- PubMed 16 Million abstracts
- MedlinePlus – health information
- OMIM – catalog of human genes and genetic disorders

Undiscovered public knowledge may have also increased by a large amount

Semantic Associations over RDF graphs

## – Discovery and Ranking

It is therefore critical to bridge the gap between unstructured and structured data by extracting entities and relationships between resulting in semantic metadata

# Outline

Motivation

Problem Description & Approach

Results

Future Work

# Problem – Extracting relationships between MeSH terms from PubMed

# Background knowledge used

UMLS – A high level schema of the biomedical domain

- 136 classes and 49 relationships
- Synonyms of all relationship – using variant lookup (tools from NLM)
- 49 relationship + their synonyms = ~350 mostly verbs

MeSH

- 22,000+ topics organized as a forest of 16 trees
- Used to query PubMed

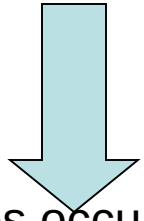
PubMed

- Over 16 million abstract
- Abstracts annotated with one or more MeSH terms

**T147—effect**
**T147—induce**
**T147—etiology**
**T147—cause**
**T147—effecting**
**T147—induced**

# Method – Parse Sentences in PubMed

[1254239-1] An excessive endogenous or exogenous stimulation by estrogen induces adenomatous hyperplasia of the endometrium.



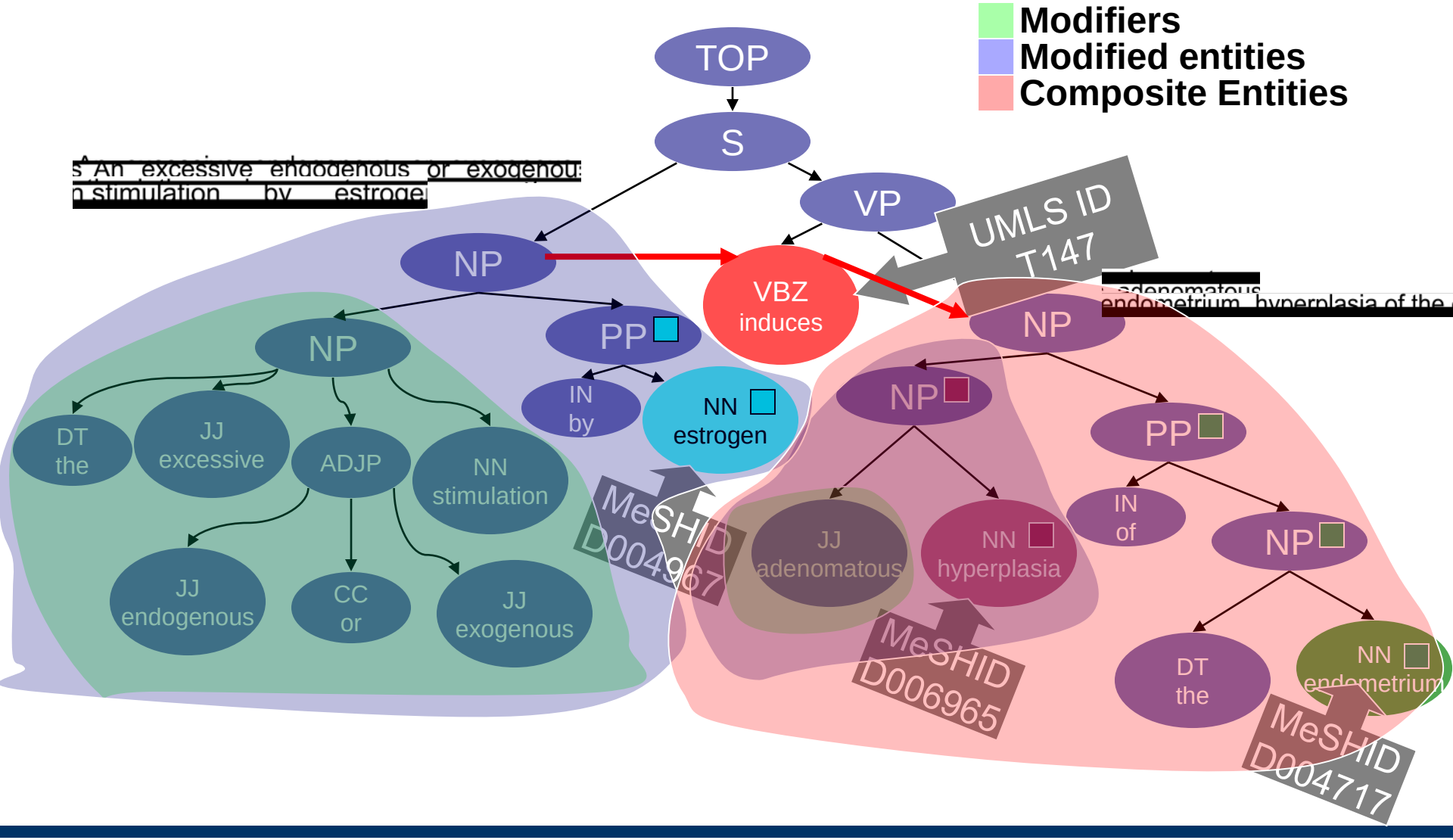
**SS-Tagger (University of Tokyo)**

**SS-Parser (University of Tokyo)**

**(TOP (S (NP (NP (DT An) (JJ excessive) (ADJP (JJ endogenous) (CC or) (JJ exogenous) ) (NN stimulation) ) (PP (IN by) (NP (NN estrogen) ) ) (VP (VBZ induces) (NP (NP (JJ adenomatous) (NN hyperplasia) ) (PP (IN of) (NP (DT the) (NN endometrium) ) ) ) ) ) )**

- Entities (MeSH terms) in sentences occur in modified forms
  - "adenomatous" modifies "hyperplasia"
  - "An excessive endogenous or exogenous stimulation" modifies "estrogen"
- Entities can also occur as composites of 2 or more other entities
  - "adenomatous hyperplasia" and "endometrium" occur as "adenomatous hyperplasia of the endometrium"

# Method – Identify entities and Relationships in Parse Tree

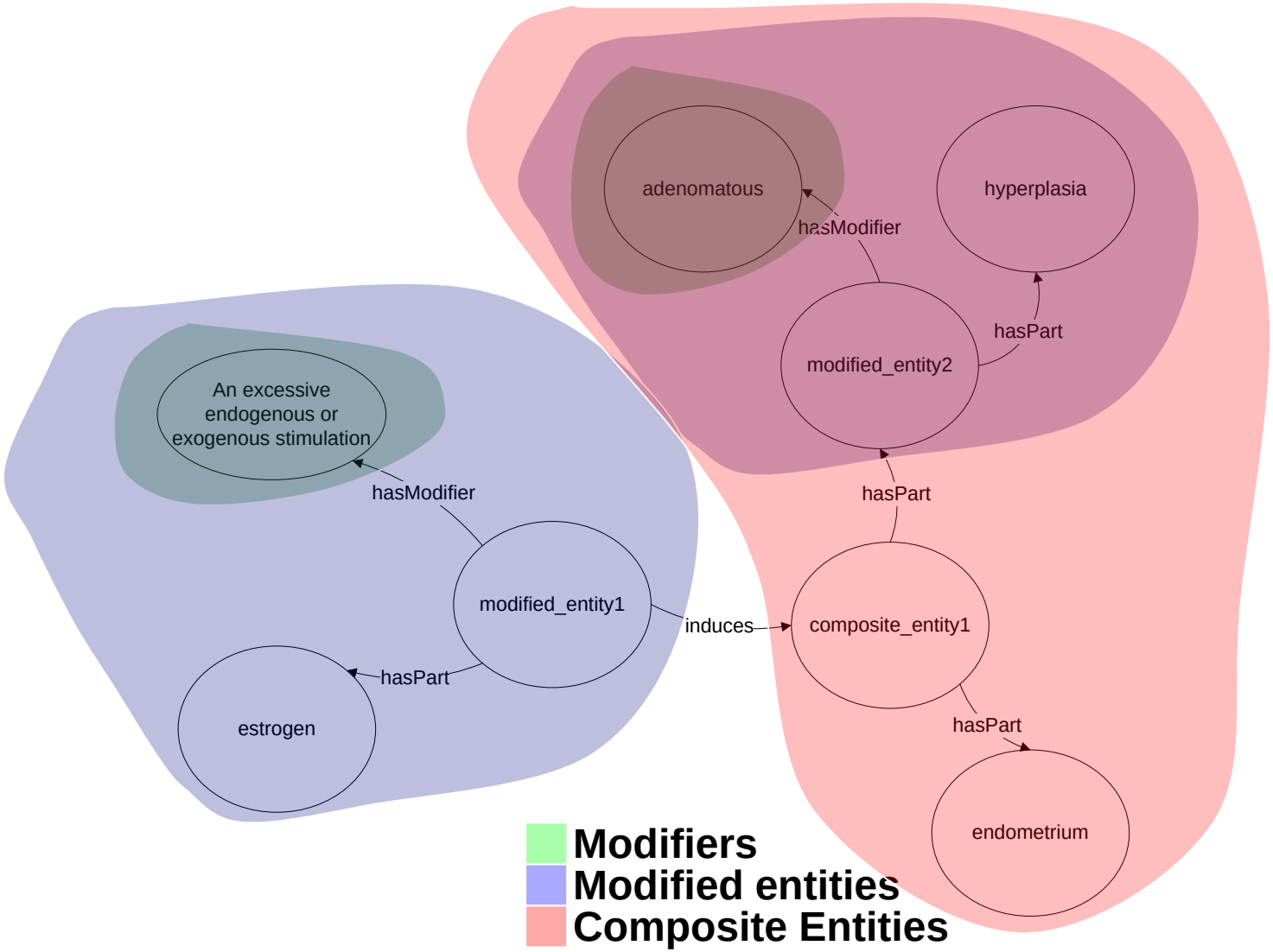


# Entities – The simple, the modified and the composite

To capture the various types of entities we define

- Simple entities as MeSH terms
- Modifiers as siblings of entities that are
  - Determiners – "Y induces no X"
  - Noun Phrases – "An excessive endogenous or exogenous stimulation"
  - Adjective phrases – "adenomatous"
  - Prepositional phrases – "M is induced by the X in the Z"
- Modified Entities as any entity that has a sibling which is a modifier
- Composite Entity as any entity that has another entity as a sibling

# Resulting RDF



# Outline

Motivation

Approach

Results

Future Work

Dataset 1

- Swanson's discoveries
  - Associations between Migraine and Magnesium [Hearst99]

    stress is associated with migraines

    stress can lead to loss of magnesium

    calcium channel blockers prevent some migraines

    magnesium is a natural calcium channel blocker

    spreading cortical depression (SCD) is implicated in some migraines

    high levels of magnesium inhibit SCD

    migraine patients have high platelet aggregability

    magnesium can suppress platelet aggregability

# Results – Creation of Dataset 1

Keywords pairs e.g. stress + migraine etc. against PubMed return PubMed abstracts that are annotated (by NLM) with both terms

8 pairs of terms in this scenario result in 8 subsets of PubMed

Semantic Metadata

- Represented in RDF
- With complex entities and relationships connecting them
- Pointers to original document and sentence
- Size
  - ~2MB RDF for Migraine Magnesium subset of PubMed

# Evaluating the Result of Extraction

Ideal method to evaluate the Extraction method

- Domain experts read a set of abstract given a set of relationship names and entities to look for
- In addition to this give them the extracted triples and entities
- For every abstract the expert judges counts the correct, incorrect and missed triples
- Measure precision and recall

# Evaluating the Result of Extraction

In the absence of a domain expert we focus of getting a feel for the utility of the extracted data

- We know the association manually discovered between Migraine and Magnesium
- We locate paths of various lengths between them and manually inspect these paths
- If the paths are indicative of the manually discovered associations the extracted data is useful

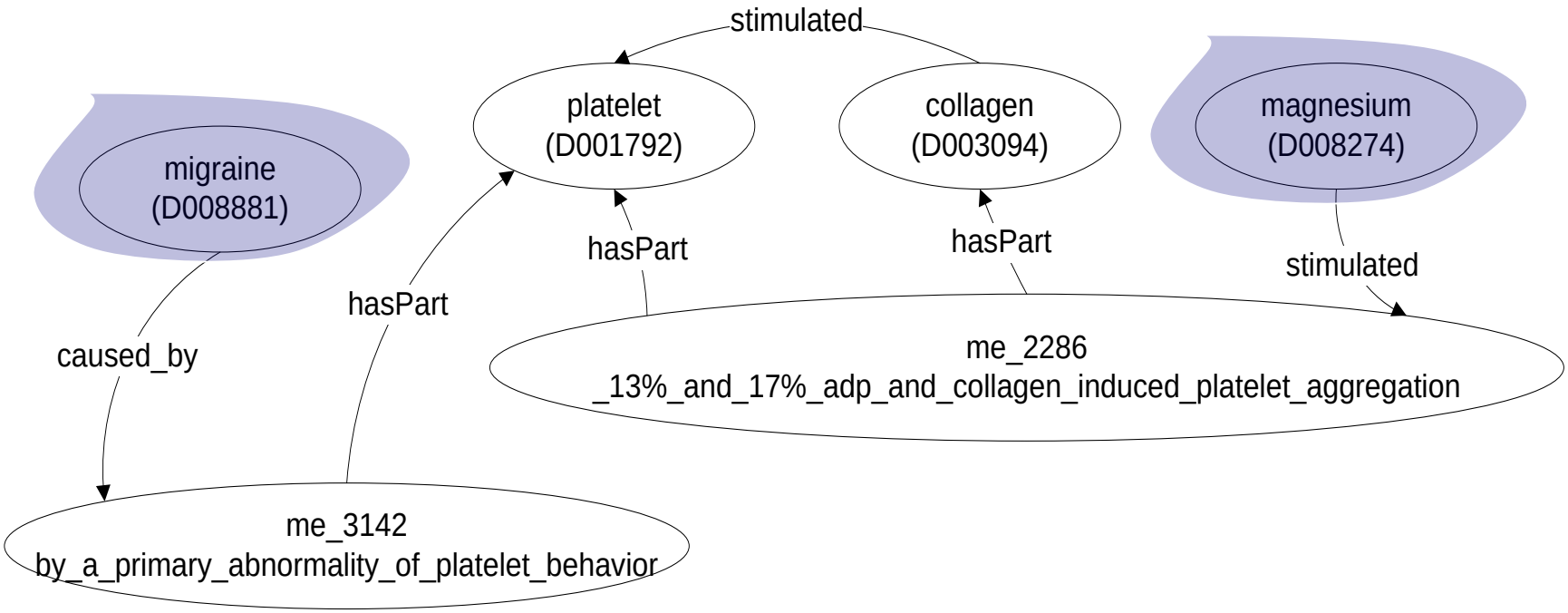
# Paths between Migraine and Magnesium

**Table 2.** Paths between Migraine and Magnesium

| Paths between Migraine and Magnesium | | | |
|---|---|---|---|
| Path length | Total Number of paths found | # of interesting paths | Max. # of named relationships in any path |
| *6* | *260* | *54* | *4* |
| *8* | *4103* | *1864* | *5* |
| *10* | *106450* | *33403* | *5* |

Paths are considered interesting if they have one or more named relationship Other than ***hasPart*** or ***hasModifiers*** in them

# An example of such a path



# Results

Dataset 2

- Neoplasm (C04)
  - For subtree of MeSH rooted at Neoplasms all topics under this subtree are used as query terms against PubMed
  - The resulting dataset contains ~500,000 PubMed abstracts
  - The extraction process run on this data returns ~150MB

Processing the tagged and parsed sentences for Dataset 2 (Neoplasm) to generate RDF took approx. 5 minutes

Stats

- 211 different named relationships found
- 500,000 instance-property-instance statements
- 260,000 instance-property-literal statements

Currently setting up to extract RDF from all of PubMed

# Outline

Motivation

Problem Description & Approach

Results

Future Work

# Future Extensions to the Extraction process

Short-term goals (1 month)

- MeSH qualifiers (blood pressure, contraindications)
- Curate and release Migraine-Magnesium RDF

Long-Term goals

- More complex structures
  - Conjunctions
  - X causes Y to inhibit Z
- Rule-action language to test new extraction rules
- Finding new terms to enrich existing vocabularies
- Perhaps ontology enrichment

# The projected future of research in Biology

From ...

Hypothesis driven “wet lab” experiments

To ...

Data-driven reduction/pruning of hypothesis space leading to new insight and possibly discovery

What challenges does this transition bring?



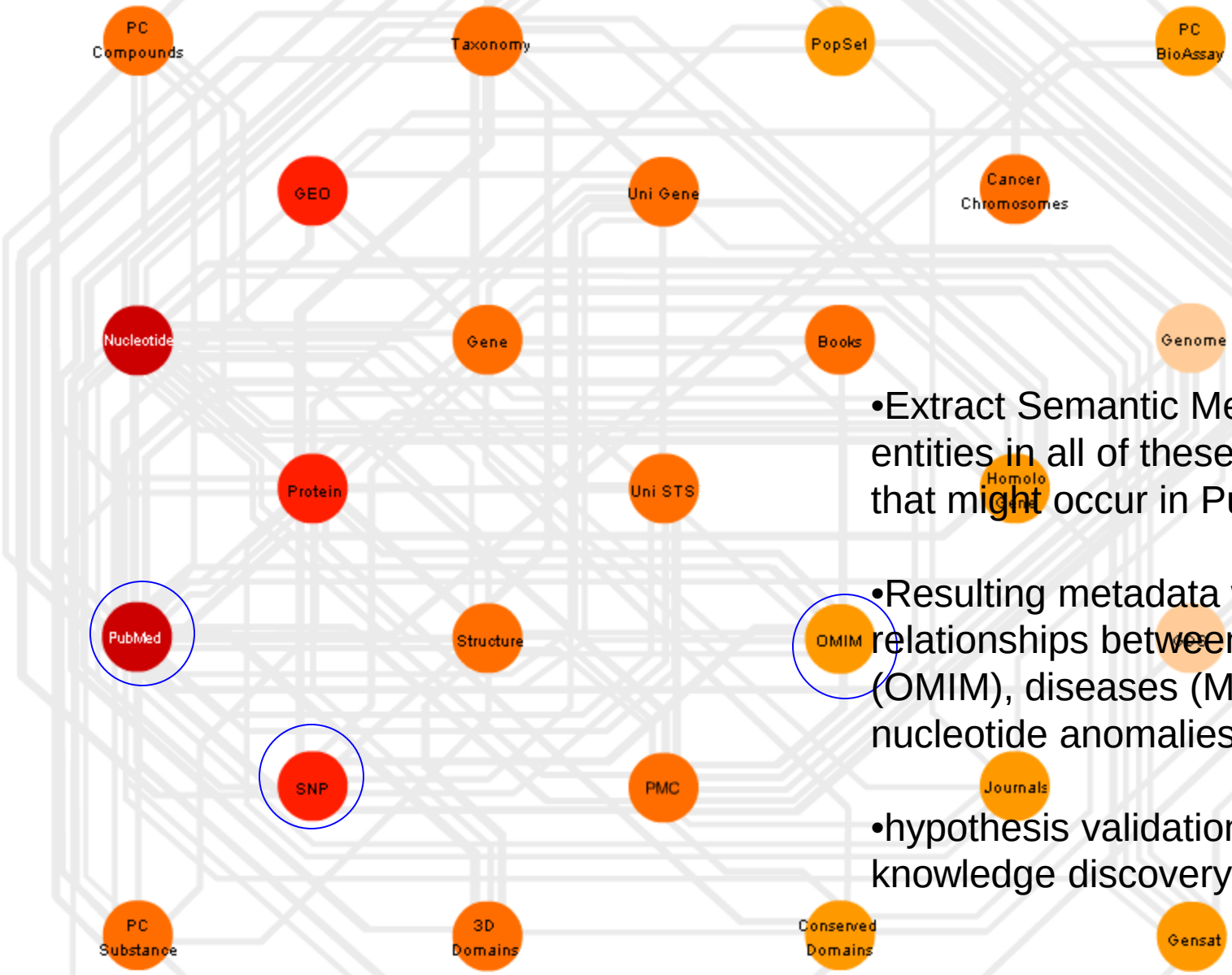
# Use of Generated Semantic Metadata

Semantic Browsing of PubMed based on named relationships between MeSH terms

Path/hypothesis based document retrieval

Knowledge discovery from literature

- Coprus-based complex relationship discovery and ranking
- Corpus-based relevant connection subgraph discovery




- Extract Semantic Metadata about entities in all of these databases that might occur in PubMed text
- Resulting metadata will contain relationships between genes (OMIM), diseases (MeSH), nucleotide anomalies (SNP)
- hypothesis validation and knowledge discovery in biology.

THANK YOU!